FoGConcept: Identifying Interpretable Movement Concepts for Freezing-of-Gait Detection from Multi-Scale IMU Signals

Anonymous authors
Paper under double-blind review

Abstract

Freezing of gait (FoG) is a disabling motor symptom of Parkinson's disease, characterized by an episodic inability to initiate or sustain effective stepping. Current inertial measurement unit (IMU)-based models can detect FoG accurately, but they fail to explain their predictions in terms of human-understandable movement concepts. A major contribution of this work is to identify nine human-understandable movement concepts encoded by a multi-scale FoG model, including gait initiation, rhythmic bilateral gait, trembling in place, shuffling steps, and pivot turns. These concepts recur across subjects, are functionally linked to the model's FoG decisions, and exhibit emergent specialization across short and long temporal contexts. Accordingly, we propose FoGConcept, a mechanism-based interpretability framework for learning movement concepts from multi-scale IMU signals. Domain experts independently recognize 88.9% of the identified concepts, while the proposed model achieves an F1 score of 94.2% for FoG detection. Our findings show that accurate FoG detection can be understood through human-recognizable movement patterns, providing new insight into what a detection model learns to distinguish FoG from non-FoG movement.

1 Introduction

Freezing of gait (FoG) is a disabling motor symptom of Parkinson's disease in which effective stepping is briefly interrupted despite the intention to move. FoG restricts mobility, increases fall risk, and substantially affects quality of life (Bloem et al., 2004; Lord et al., 2020). Wearable inertial measurement units (IMUs) enable continuous gait monitoring and have supported increasingly accurate machine-learning (ML) methods for automatic FoG detection (Salomon et al., 2024; Koltermann et al., 2024; Chi et al., 2026). However, predictive accuracy alone does not explain how these models make their decisions or what they have learned about human movement.

Existing interpretation methods for FoG detection models primarily expose low-level predictive evidence, such as an individual sensor channel, signal peak, or kinematic attributes (Filtjens et al., 2021; Al-Adhaileh et al., 2025). These methods show which inputs influence a prediction, but not what internal representations support the model's decision. We therefore ask: Can the internal representations supporting FoG detection be interpreted in a human-understandable manner?

To investigate this question, we introduce FoGConcept, a mechanism-based interpretability framework for learning movement concepts from multi-scale IMU signals. Specifically, we first build a multi-scale model that encodes bilateral IMU signals over short, medium, and long temporal contexts and fuses their representations for detection. A shared sparse autoencoder (SAE) then decomposes token representations across scales into sparse features. We identify human-understandable concepts from latent features through representation steering and movement-pattern characterization, and test their reproducibility through cross-subject validation. The resulting concepts are further validated by blinded domain experts.

Our analysis reveals that the model's decision-relevant features can be explained through high-level, human-understandable movement patterns. We call a recurring, human-recognizable movement pattern associated with a model-relevant feature a movement concept. We identify nine such concepts spanning localized unilateral events like single-foot strikes, locomotor transitions like the start and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

end of walking, sustained normal gaits, rotational coordination, PD-associated abnormal gaits like shuffling, and FoG-specific patterns such as the start of freezing and trembling in place.

Fig. 1 previews two examples of interpretable movement concepts. L6 represents *Rhythmic Gait Cycles*, with regular acceleration peaks alternating between the two feet across successive steps. L1554 represents *Shuffling Steps*, where a person moves forward with very small steps and drags their feet along the ground instead of lifting them fully (Schaafsma et al., 2003). Here, irregular stepping attempts are evident in the gyroscope signals but produce only weak acceleration responses, consistent with limited effective progression.

Beyond their movement semantics, the discovered concepts exhibit emergent specialization across temporal contexts. Localized patterns such as the start of freezing (Fig. 3b) preferentially emerge at shorter contexts, whereas sustained or multi-cycle patterns such as rhythmic gait emerge at longer ones. The associated movement patterns also recur across held-out subjects, remain distinct from one another, and are independently recognizable by domain experts.

In summary, our contributions are:

- We contribute human-understandable movement concepts underlying FoG detection, including gait initiation, single foot strikes, bilateral foot oscillation, and find emergent specialization across temporal contexts.
- We propose FoGConcept, the first mechanism-based interpretable framework that identifies aforementioned human-understandable concepts from multi-scale IMU data.
- We evaluate FoGConcept on real-world FoG data. Domain experts recognize 88.9% of the discovered concepts, and the underlying multi-scale model achieves 94.2% F1 for FoG detection.

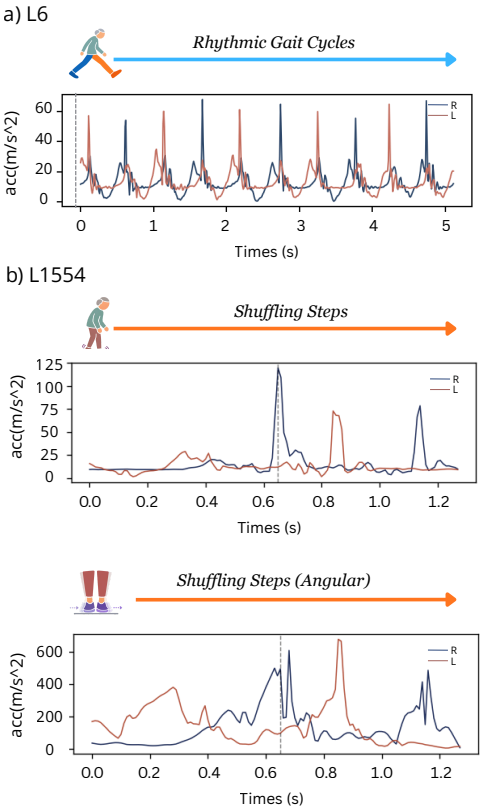


Figure 1: **Examples of movement concepts.** Each panel shows a high-activation window from bilateral ankle IMUs. Red and blue denote the left and right foot, respectively. Acceleration shows the three-axis magnitude in $\mathrm{m/s^2}$; angular velocity shows the three-axis magnitude in $^\circ/\mathrm{s}$. (a) L6 captures rhythmic bilateral gait cycles with regular alternating foot activity. (b) L1554 captures shuffling steps, where repeated stepping attempts are evident in angular velocity while the corresponding acceleration is small and irregular.

2 Related Work

FoG Detection and Interpretation. Wearable IMUs are widely used for FoG detection. Earlier methods combine hand-crafted features with conventional classifiers (Moore et al., 2008), while recent CNN-, RNN-, transformer-, and hybrid-based approaches learn temporal gait representations directly from wearable signals (Delgado-Terán et al., 2025; Koltermann et al., 2024; Chi et al., 2026). Although these models achieve strong predictive performance, they're black boxes and remain difficult to interpret.

Existing interpretation of learning-based FoG models is predominantly input-level. Attention mechanisms highlight salient portions of sensor sequences (Al-Adhaileh et al., 2025), while layer-wise relevance propagation assigns relevance to input dimensions associated with FoG predictions (Filtjens et al., 2021). These methods identify influential input evidence but do not directly characterize

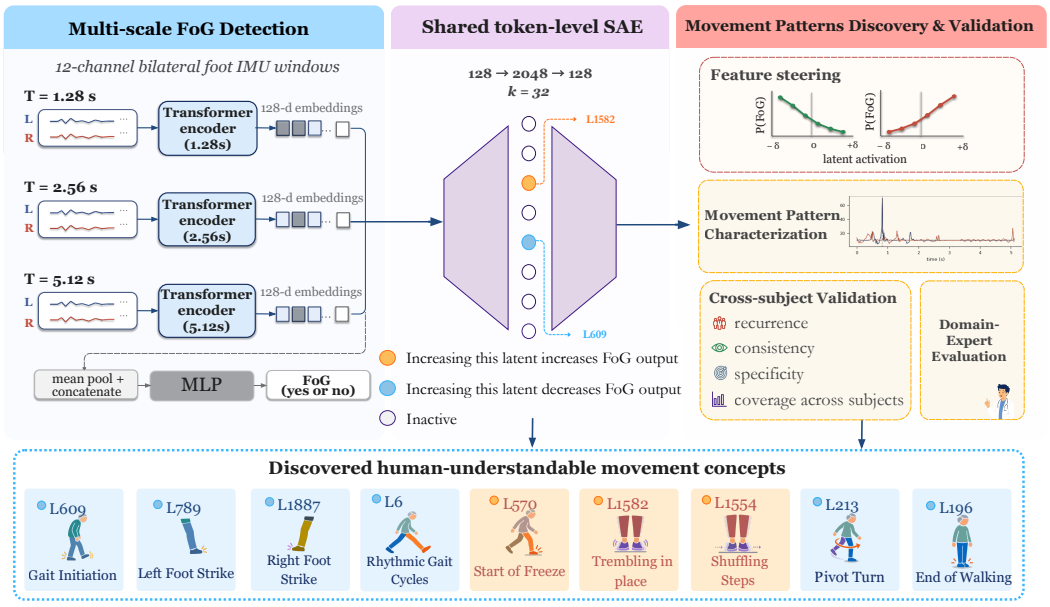


Figure 2: **Overview of FoGConcept.** Bilateral ankle-IMU signals are encoded at three temporal scales and fused for FoG detection. A shared token-level SAE decomposes the encoder representations into sparse features. Model-relevant features are identified through steering, characterized from their high-activating IMU windows, and validated across subjects. The resulting movement patterns are further assessed by domain experts.

the movement patterns encoded in internal representations. Our work instead asks whether functionally relevant internal features correspond to human-recognizable movement patterns.

Temporal context is also important for FoG detection. Prior work either evaluates different window lengths as separate model settings or combines features computed from multiple temporal windows (Kwon et al., 2014; Cohen et al., 2024). In contrast, our model jointly processes raw IMU signals at multiple input durations, learns a separate representation at each temporal scale, and fuses these representations for a single FoG prediction. This design integrates short- and long-term gait dynamics within one model, and also enables us to examine how internal movement concepts specialize across temporal contexts.

Sparse Autoencoders for Interpretability. Sparse autoencoders (SAEs) decompose dense neural representations into sparse latent features and have been used to identify interpretable internal structure and test feature function through interventions (Cunningham et al., 2023; Gao et al., 2024). SAE-based interpretability has expanded from language and vision (Lim et al., 2025)to speech (Du et al., 2026; Koriagin et al., 2026) and time-series models (Kalnāre et al., 2025; Oublal et al., 2026; Mishra, 2026). These studies demonstrate that sparse feature analysis can be applied to temporal representations, but they do not examine physiological movement encoded by wearable IMUs.

The closest physiological precedent applies SAEs to EEG foundation-model representations and evaluates latent features against predefined clinical concepts, together with concept steering and spectral analysis (Lehn-Schiøler et al., 2026). In contrast, we do not begin with a predefined movement-concept vocabulary. Model-relevant latents are selected by their functional effect and subsequently evaluated for cross-subject consistency and blinded expert recognition. To our knowledge, this is the first SAE-based analysis of wearable FoG representations based on human-understandable movement concepts.

3 Methodology

3.1 Overview of FoGConcept

Fig. 2 summarizes the proposed multi-scale interpretable framework, FoGConcept. A multi-scale FoG detection model encodes bilateral ankle-IMU signals at three temporal scales and combines the

resulting representations for FoG detection. We then decompose its token-level representations with a shared SAE and analyze the learned SAE features through steering, movement-pattern characterization, and cross-subject validation.

The FoG model takes synchronized signals from two ankle-mounted IMUs. At each timestep, each IMU records three-axis linear acceleration and three-axis angular velocity from the left and right feet. For each prediction anchor, we extract three aligned windows $\boldsymbol{X}_s \in \mathbb{R}^{T_s \times 12}$ with $T_s \in \{128, 256, 512\}$ samples, corresponding to 1.28, 2.56, and 5.12s. Each window ends at the prediction anchor.

Each scale uses the Transformer configuration and masked-reconstruction pretraining procedure of FM-FoG (Chi et al., 2026), with only the input length changed across scales. The three encoders are independently parameterized. For scale s, the encoder produces $\boldsymbol{H}_s = [\boldsymbol{h}_{s,1}, \ldots, \boldsymbol{h}_{s,T_s}]$, where each token $\boldsymbol{h}_{s,t} \in \mathbb{R}^{128}$. The sequence is mean-pooled into a 128-dimensional representation. The three pooled representations are then concatenated into a 384-dimensional vector and passed to an MLP classifier to obtain $P(\mathrm{FoG})$. We use mean pooling so that token positions contribute uniformly to the pooled representation, avoiding learned position-dependent weighting in the subsequent intervention analysis. The trained FoG model is frozen for all interpretability analyses.

For interpretability, we train a shared TopK SAE (Makhzani & Frey, 2014; Gao et al., 2024) on the token representations $\boldsymbol{h}_{s,t}$ before pooling and cross-scale fusion. At this stage, each token still corresponds to one input timestep, allowing SAE activations to be traced back to localized regions of the IMU signal. A single SAE is trained on tokens from all three encoder streams.

For a token $\boldsymbol{h} \in \mathbb{R}^d$ with $d = 128$, the SAE encoder produces a sparse feature-activation vector $\boldsymbol{z} \in \mathbb{R}^F$:

$$\boldsymbol{z} = \mathrm{TopK}_k(\mathrm{ReLU}(\boldsymbol{W}_{\mathrm{enc}}(\boldsymbol{h} - \boldsymbol{b}_{\mathrm{pre}}) + \boldsymbol{b}_{\mathrm{enc}})). \tag{1}$$

Here, $\boldsymbol{W}_{\mathrm{enc}} \in \mathbb{R}^{F \times d}$ and $\boldsymbol{b}_{\mathrm{enc}} \in \mathbb{R}^F$ are learned encoder parameters, and $\boldsymbol{b}_{\mathrm{pre}} \in \mathbb{R}^d$ is a learned centering bias. The encoder centers the token representation, projects it into an F-dimensional feature space, and retains only the k strongest positive feature responses.

The decoder maps the sparse representation back to the original token space:

$$\hat{\boldsymbol{h}} = \boldsymbol{W}_{\mathrm{dec}}\boldsymbol{z} + \boldsymbol{b}_{\mathrm{pre}} = \boldsymbol{b}_{\mathrm{pre}} + \sum_{f=1}^{F} z_f \boldsymbol{d}_f, \tag{2}$$

where $\boldsymbol{W}_{\mathrm{dec}} \in \mathbb{R}^{d \times F}$ is the learned decoder matrix and $\boldsymbol{d}_f$ denotes its fth column. Thus, the reconstructed token is expressed as a weighted combination of learned decoder directions: z_f specifies how strongly SAE feature f is active; $\boldsymbol{d}_f$ specifies the corresponding direction in the model's representation space. The SAE minimizes reconstruction mean squared error, with sparsity imposed directly by the TopK operation, and is frozen after training.

3.2 Movement Pattern Discovery & Validation

A key challenge for the proposed framework is identifying human-understandable concepts from latent features. To address this challenge, we identify movement concepts through a three-stage process. First, feature steering selects SAE features that influence the model's FoG prediction. Second, we characterize the movement pattern associated with each selected feature from its high-activation IMU windows. Third, we test whether these patterns recur and remain distinct across subjects in an independent raw-signal representation.

Feature Steering. SAE decoder directions provide a direct interface for intervening on learned features and measuring their effect on downstream model behavior (Du et al., 2026). We therefore use steering to identify model-relevant features before inspecting their movement semantics. For feature f, let $z_f(\boldsymbol{h}_{s,t})$ denote its activation on token $\boldsymbol{h}_{s,t}$ and $\boldsymbol{d}_f$ its decoder direction. We intervene only at token positions where the feature is active:

$$\tilde{\boldsymbol{h}}_{s,t}^{(f)}(\alpha) = \begin{cases} \boldsymbol{h}_{s,t} + \alpha \boldsymbol{d}_f, & z_f(\boldsymbol{h}_{s,t}) > 0, \\ \boldsymbol{h}_{s,t}, & z_f(\boldsymbol{h}_{s,t}) = 0. \end{cases} \tag{3}$$

The intervention therefore displaces an active token along the representation-space direction associated with feature f, while leaving inactive tokens unchanged. Positive α shifts the representation

in the feature direction, negative α shifts it in the opposite direction, and $\alpha = 0$ recovers the original representation. The edited tokens are propagated through the unchanged pooling, fusion, and classification layers.

For each feature, we evaluate a fixed set of feature-active host windows over a feature-specific range of α. Let $\bar{p}_f(\alpha)$ denote the mean FoG probability across these windows after steering by α. We summarize the feature's functional effect as

$$\Delta_f = \max_{\alpha \in \mathcal{A}_f} \bar{p}_f(\alpha) - \min_{\alpha \in \mathcal{A}_f} \bar{p}_f(\alpha), \tag{4}$$

where $\mathcal{A}_f$ is the steering grid. Thus, Δ_f measures how much the detector output changes when the representation is moved along feature f's decoder direction. Features whose effect exceeds a prespecified threshold are retained as model-relevant. Feature selection is completed before movement characterization.

Movement Pattern Characterization. Because SAE activations are defined at the token level, we first aggregate them to obtain one activation score per input window. For feature f and window w at scale s, we define

$$z_f(w) = \max_t z_f(\boldsymbol{h}_{s,t}). \tag{5}$$

We use the maximum response because the three temporal scales contain different numbers of tokens; unlike a token average, it does not dilute a localized activation in longer sequences. Windows are ranked by $z_f(w)$ to identify the examples in which feature f is most strongly expressed.

Following the common SAE practice of interpreting features through maximally activating examples (Bricken et al., 2023; Lim et al., 2025), we inspect the highest-ranking windows. We examine the corresponding bilateral acceleration and angular velocity magnitude traces, together with their frequency content, to identify recurring movement patterns. When the high-activation examples exhibit a consistent, human-understandable movement pattern, we refer to that interpretation as the movement concept associated with feature f.

Cross-Subject Validation. High-activation examples alone may reflect subject-specific structure, so we test whether each characterized feature (i) recurs in recurs in subjects held out from movement characterization, (ii) preserves its characteristic signal form, and (iii) remains distinct from other features. Each window is represented by four magnitude trajectories: left- and right-foot acceleration and angular velocity. Each trajectory is resampled to 32 points, and the four trajectories are concatenated into a fixed 128-dimensional descriptor $\boldsymbol{\phi}(w) \in \mathbb{R}^{128}$. For feature f, windows are labeled active ($z_f(w) > 0$) or inactive, with inactive controls matched by temporal scale; all probes and prototypes are constructed on characterization subjects and applied without refitting to evaluation subjects.

Recurrence trains a logistic probe $\boldsymbol{\phi}(w) \to \{\text{active}, \text{inactive}\}$ and computes held-out ROC-AUC from the active/inactive labels and predicted active-class scores, testing whether feature-associated signal structure reappears across subjects. *Prototype consistency* uses no classifier: active characterization windows are averaged into a fixed prototype, each evaluation window is scored by cosine similarity to it, and ROC-AUC tests whether active windows are more prototype-like than inactive controls. *Specificity* trains pairwise probes $\boldsymbol{\phi}(w) \to \{f_i, f_j\}$ and evaluates held-out ROC-AUC, testing whether different feature-associated patterns remain distinguishable. Overlapping windows are assigned to the feature with stronger activation. Full definitions and implementation details are provided in Appendix D.

4 Evaluation

4.1 Experimental Settings

Dataset and FoG detection setup. We use the VCU FoG-IMU dataset (Chi et al., 2026), containing bilateral ankle-IMU recordings from 23 participants with Parkinson's disease (237.9 min; 528 annotated FoG episodes) sampled at 100Hz (128 samples for a 1.28s window). FoG detection is evaluated using 23-fold leave-one-subject-out (LOSO) cross-validation. For each prediction anchor, the 2.56s window defines the label: a sample is positive if FoG occurs within the final 30% of the

window and negative otherwise (Koltermann et al., 2024; Chi et al., 2026). The aligned 1.28, 2.56, and 5.12s windows end at the same prediction anchor and are fused to predict this common label. We re-run all external baselines with 2.56s inputs using the same preprocessing, labeling rule, and LOSO folds, and additionally evaluate a single-scale 2.56s version of our model. Implementation details are provided in Appendix A.

Interpretability setup. For interpretability, the multi-scale model and shared SAE are trained on all 23 participants and frozen before analysis. We use $d_{\mathrm{SAE}} = 2048$ and $k = 32$; sensitivity to SAE configuration is reported in Appendix E. The 10/13 subject split is used only for post-hoc concept characterization and validation: 10 participants form the characterization group and 13 disjoint participants form the evaluation group. For steering, host windows must activate the target feature and have $P(\mathrm{FoG}) \in [0.15, 0.85]$. For feature f, we sweep $\alpha \in [-3m_f, 3m_f]$, where m_f is its median nonzero activation, and retain features with $\Delta_f \geq 0.10$.

Metrics. FoG detection is evaluated using F1, accuracy, precision, and recall, with F1 as the primary metric. Cross-subject recurrence, prototype consistency, and pairwise specificity are evaluated using ROC-AUC on the disjoint evaluation subjects; subject coverage is reported descriptively. SAE representation fidelity and configuration sensitivity are reported in Appendix E. Because prior interpretable FoG methods provide input-level attributions rather than discovered internal concepts, there is no directly comparable concept-discovery baseline; we therefore evaluate interpretability through functional, cross-subject, and expert validation.

4.2 Human-Understandable Movement Concepts

We first examine the movement patterns represented by the model relevant SAE features. Fig. 1 and Fig. 3 collectively show six representative concepts identified from their high-activation windows (complete atlas in Appendix B). Each trace shows the Euclidean magnitude across the three sensor axes. Acceleration is reported in $\mathrm{m/s^2}$ and angular velocity in $^\circ/\mathrm{s}$; red and blue denote the left and right foot, respectively. Because acceleration magnitude includes gravity, a stationary foot appears as a nearly constant trace around $9.8\,\mathrm{m/s^2}$. We show the modality that most clearly exposes each pattern, and both modalities when their relationship is informative. Vertical dashed markers indicate the temporal location of maximum feature activation where shown.

These concepts do not all represent FoG itself; rather, they span ordinary gait, PD-associated gait abnormalities, transitions into FoG, and recognized FoG phenomena that contribute to the model's FoG decisions. First, they capture distinct structure within ordinary gait. L609 (*Gait Initiation*; Fig. 3a) shows both feet remaining close to a stationary acceleration baseline before the feature activates. At the activation point, the traces depart from this baseline and bilateral foot movement begins, followed by repeated stepping activity. The feature therefore marks the transition from standing to walking. L6 (*Rhythmic Gait Cycles*; Fig. 1a) captures the sustained normal walking pattern. Across the window, acceleration peaks recur at a regular cadence and alternate between the left and right feet, producing the bilateral cycles characteristic of continuous walking. L609 and L6 separate the onset of locomotion from the rhythmic organization of established gait.

L570 (*Start of Freeze*; Fig. 3b) captures the transition from an effective step into FoG. The window contains a single prominent right-foot acceleration spike, corresponding to the final effective step. Maximum feature activation occurs on the falling edge of this spike, precisely as the step terminates. For most of the following interval, both feet return close to near-stationary, indicating little foot movement corresponding with the annotated FoG period. L570 therefore localizes the boundary at which effective stepping ends and freezing begins. This transition emerges without explicit supervision for gait phase or FoG-onset timing.

L1582 (*Trembling in Place*; Fig. 3c) captures a different organization of movement around FoG. The acceleration traces contain dense, rapid fluctuations from both feet, punctuated by larger transient peaks. Unlike L6, however, these fluctuations do not recur at a stable cadence or form a consistent alternating left-right sequence. A near-stationary interval is also visible around the middle of the example during freezing, with irregular movement appearing around it. Thus, the lower limbs can remain highly active without forming the coordinated gait cycles required for effective walking. This pattern is consistent with trembling-in-place FoG, characterized by rapid leg movements with little or no effective forward progression (Nutt et al., 2011). The contrast between L6 and L1582

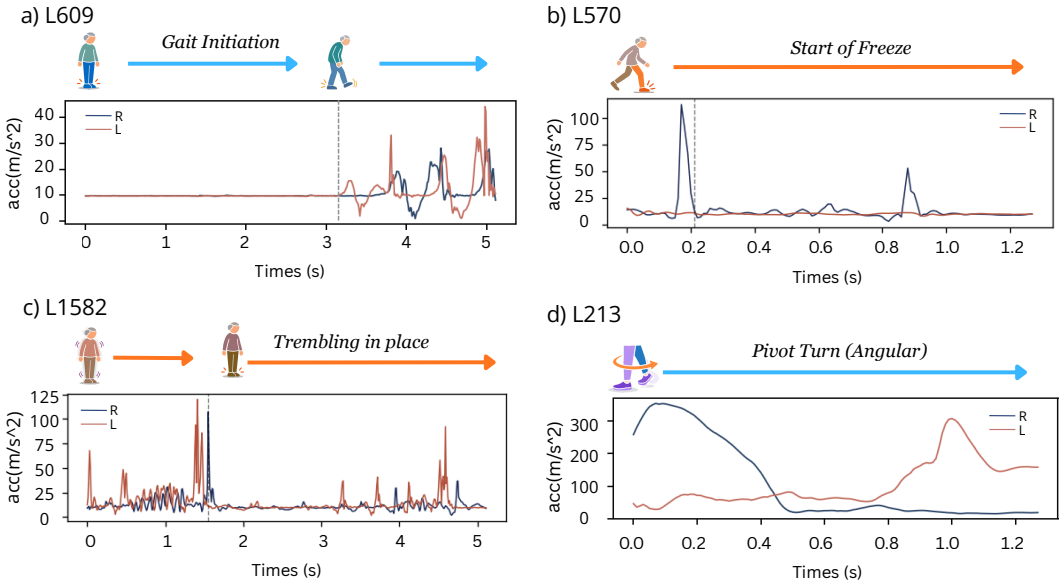


Figure 3: **Representative human-understandable movement concepts.** Red and blue denote the left and right foot, respectively. Acceleration magnitude is shown for (a) L609 *Gait Initiation*, from standing still to walking, (b) L570 *Start of Freeze*, the transition from a step into FoG and (c) L1582 *Trembling in Place*, represented by rapid lower-limb movement without stable gait cycles; angular-velocity magnitude is shown for (d) L213 *Pivot Turn*, the rotational movement of the two feet when turning around. Vertical dashed lines indicate the temporal location of maximum feature activation.

shows that the representation distinguishes organized rhythmic gait from high-frequency movement that does not resolve into effective stepping.

L1554 (*Shuffling Steps*; Fig. 1b) represents another form of disrupted gait and requires the combined analysis of both acceleration and angular velocity. In PD, a shuffling gait is characterized by very short, low-clearance steps that produce limited forward progression while stepping continues (Nutt et al., 2011). This behavior is reflected in the two IMU modalities. The acceleration traces remain relatively close to stationary and lack the regular, alternating peaks seen during normal walking, indicating weak and irregular translational foot dynamics. In contrast, the gyroscope shows attempted angular motion that alternates between the two feet, demonstrating that the feet continue to rotate and try to participate in a stepping sequence. In several intervals, substantial rotation (angular excursion) of one foot is accompanied by only a small acceleration response. These signals indicate attempts in stepping with limited effective progression, consistent with the short, poorly advancing steps of shuffling.

L1554 and L1582 therefore represent clinically distinct movement patterns: shuffling is a PD-associated gait abnormality, whereas trembling in place is a recognized FoG phenotype. Shuffling preserves bilateral walking attempts, most clearly visible in angular velocity, but the associated acceleration is weak and irregular, whereas trembling in place contains rapid lower-limb activity without stable gait cycles. The representation thus separates distinct organizations of impaired movement, and does not collapse different abnormal PD gait manifestations into a generic abnormal-motion feature.

Finally, L213 (*Pivot Turn*; Fig. 3d) captures bilateral coordination in rotational motion. The right-foot gyroscope first forms a large, smooth, bell-like excursion in angular-velocity while the left foot remains stable. As the right-foot excursion returns toward baseline, the roles reverse: the left foot develops a similarly broad rotational excursion while right-foot angular motion remains low. These two sequential rotational bouts are consistent with one foot rotating while the contralateral foot acts as support, followed by an exchange of roles as the turn progresses.

Beyond their movement semantics, the recovered features exhibit emergent temporal-scale organization. Although the shared SAE receives no explicit scale identifier, five of the nine concepts have their highest scale-normalized firing rate in the 1.28s stream, while four are dominated by the

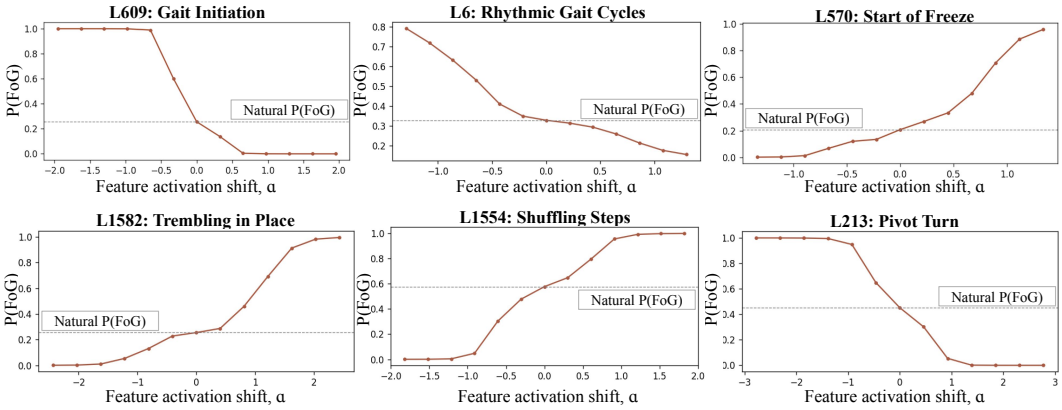


Figure 4: **Representative feature-steering responses for movement concepts.** Each curve shows the mean predicted $P(\text{FoG})$ as the activation of one SAE feature is shifted by α along its decoder direction. The dashed line marks the prediction at the feature's natural activation, with $\alpha = 0$ corresponding to the unmodified representation. Positive steering of L570, L1582, and L1554 increases the FoG prediction, whereas positive steering of L609, L6, and L213 decreases it.

5.12s stream (Appendix C). The short-context concepts include Left and Right Foot Strike, Start of Freeze, Shuffling Steps, and Pivot Turn, whereas Gait Initiation, Rhythmic Gait Cycles, Trembling in Place, and End of Walking prefer the long context. This organization aligns with movement duration: localized events and transitions preferentially occur at shorter contexts, whereas sustained or multi-cycle patterns occur at longer contexts.

4.3 Evidence for Functional Relevance and Interpretability

We next test whether the recovered movement concepts are functionally involved in the model's FoG decisions and validate that their associated signal patterns are not subject-specific.

Functional relevance. Fig. 4 shows steering effects for six representative concepts (complete steering effects in Appendix C). Positive steering of L570, L1582, and L1554 increases $P(\text{FoG})$, whereas steering L609, L6, and L213 decreases it. Several interventions move the prediction across a large fraction of its $[0, 1]$ range, showing that individual sparse features can have substantial effects on the frozen classifier. Steering directions show features for both FoG episodes and normal gaits.

Random-intervention control. We repeat steering with 1,000 norm-matched random directions while preserving the same host windows, token positions, and intervention magnitudes, and steering along the concepts' directions produce significantly larger effects than the random baseline (full details in Appendix C.1), indicating that the steering effects are direction-specific, not a generic consequence of perturbing the representation.

Cross-subject generality and distinctness. The discovered movement patterns generalize beyond the subjects used for their characterization. Recurrence AUC ranges from 0.90 to 0.99 (median 0.96). Thus, for the median concept, a feature-active window receives a higher recurrence score than a matched inactive control in approximately 96% of positive–negative comparisons, showing that feature-associated raw-signal structure remains strongly distinguishable from inactive controls. Prototype-consistency AUC ranges from 0.88 to 1.00 (median 0.95), showing that held-out active windows also remain more similar than controls to the fixed pattern estimated from characterization subjects. The concepts occur in 15–23 participants (median 19), not confined to isolated individuals.

The concepts also remain distinct from one another: pairwise specificity has a median held-out AUC of 0.88, including 0.92 among same-scale pairs and 0.86 across scales. These results show that the recovered concepts correspond to movement patterns that recur across individuals, preserve their characteristic form, and remain distinguishable from other discovered patterns. Full per-concept and pairwise results are provided in Appendix D.

4.4 FoG Detection Performance

Table 1 shows that the multi-scale model achieves the best FoG detection performance, with an F1 of 94.2%. This exceeds the strongest baseline, the best-performing contest submission at 88.7%, by 5.5%. The model also achieves 94.3% accuracy, precision, and recall, indicating balanced performance across metrics.

The single-scale model achieves 92.0% F1, outperforming every external baseline. Jointly modeling the aligned 1.28, 2.56, and 5.12s contexts further improves F1 by 2.2 points. These results show that the multi-scale formulation provides predictive information beyond the 2.56s representation alone.

The movement concepts analyzed in this work therefore arise from the internal representations of a high-performing FoG model. The multi-scale gain further establishes that information across temporal contexts is useful for the detection task, complementing our analysis of how model-relevant movement features are organized across temporal scales.

Table 1: **FoG detection LOSO performance.** All baselines are re-run with 2.56s inputs under the common preprocessing, labeling, and LOSO protocol. The best external result is underlined and the best overall result is bold.

Method	F_1 (%) ↑
Dual-Level (Chen et al., 2025)	65.1
Huang et al. (Huang et al., 2024)	80.4
Borzì et al. (Borzì et al., 2025)	81.8
Delgado-Terán et al. (Delgado-Terán et al., 2025)	84.0
Xia et al. (Xia et al., 2024)	84.5
MCT-Net (Wang et al., 2024)	86.6
Abbasi et al. (Abbasi & Rezaee, 2025)	87.2
Al-Adhaileh et al. (Al-Adhaileh et al., 2025)	87.3
Gait-Guard (Koltermann et al., 2024)	88.2
Contest (best) (Salomon et al., 2024)	<u>88.7</u>
Ours, single-scale (2.56 s)	92.0
Ours, multi-scale	**94.2**

4.5 Domain-Expert Evaluation

We further assess whether the movement concepts are independently recognizable to domain experts. Three experts with a combined 20+ years of experience in Parkinson's disease, FoG, and gait analysis independently reviewed the activation examples associated with the nine concepts and described the recurring movement, without access to one another's responses. A concept is considered recognized when at least two of the three independent descriptions identify the same underlying movement.

Experts recognized 8 of 9 concepts (88.9%), including five with unanimous agreement. To test whether this level of agreement could arise from chance label alignment, we independently permuted the expert concept assignments and recomputed majority agreement. Experts agreement was significantly greater than the random-label distribution (permutation test, $p < 0.001$). The free-text responses also recovered characteristic movements. For example, L609 was described as "standing up and walking," corresponding to Gait Initiation, while L213 was described as "turning in a circle," same as Pivot Turn.

After the independent review, differences in terminology or movement specificity were resolved by reviewing the corresponding source videos and FoG annotations, and with discussion, producing the final consensus labels reported in Fig. 2.

5 Conclusion and Limitations

We identify human-understandable movement concepts underlying a multi-scale FoG model, spanning localized events and sustained gait patterns. Steering, cross-subject validation, and expert evaluation show that these concepts are functionally linked to FoG predictions, recur across individuals, and are human-recognizable. However, some limitations remain to be addressed. As training SAEs on large-scale data requires non-trivial computational resources, our analysis is limited to a single model, dataset, and a small expert cohort. The discovered concepts are not necessarily exhaustive, as we focus on representative, clearly observable features, not systematically characterizing the full latent space. Finally, steering operates only in representation space since there is no reliable generative decoder for IMUs. Future work should test concept generality across models and datasets, and develop signal-space counterfactuals.

Ethics Statement

The wearable data used in this study were collected under an Institutional Review Board (IRB)-approved human-subject research protocol with informed consent and were analyzed in de-identified form. Our analyses are intended to improve understanding of learned representations in freezing-of-gait models; the discovered movement concepts should not be interpreted as independently validated clinical biomarkers or diagnostic categories.

Reproducibility Statement

Sections 3.1–3.2 specify the multi-scale model, SAE, feature-steering, movement-characterization, and cross-subject validation procedures; Sec. 4.1 describes the dataset splits, labeling protocol, steering settings, and evaluation metrics. Appendices A–E provide implementation details, the complete concept atlas and steering analysis, cross-subject validation definitions and results, and SAE reconstruction fidelity.

AI Use Statement

Generative AI tools were used to assist with language editing and presentation refinement. All experiments and analyses were conducted by the authors. All AI-assisted content was reviewed and verified by the authors.

References

Sara Abbasi and Khosro Rezaee. Deep learning-based prediction of freezing of gait in parkinson's disease with the ensemble channel selection approach. *Brain and Behavior*, 15(1):e70206, 2025. doi: 10.1002/brb3.70206. URL https://doi.org/10.1002/brb3.70206.

M. H. Al-Adhaileh, A. Wadood, T. H. H. Aldhyani, S. Khan, M. I. Uddin, and A. H. Al-Nefaie. Deep learning techniques for detecting freezing of gait episodes in Parkinson's disease using wearable sensors. *Frontiers in Physiology*, 16:1581699, 2025. doi: 10.3389/fphys.2025.1581699.

B. R. Bloem, J. M. Hausdorff, J. E. Visser, and N. Giladi. Falls and freezing of gait in Parkinson's disease: A review of two interconnected, episodic phenomena. *Movement Disorders*, 19(8):871–884, 2004. doi: 10.1002/mds.20115.

Luigi Borzì, Luis Sigcha, Farshad Firouzi, Gabriella Olmo, and Florenc Demrozi. Edge-based freezing of gait recognition in Parkinson's disease. *Computers and Electrical Engineering*, 127: 110530, 2025. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2025.110530.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Qianhong Chen, Zhonglue Chen, Fei Zhang, Shengdi Chen, Kang Ren, and Nannan Lu. Dual-level freezing of gait recognition. *IEEE Sensors Journal*, pp. 1–1, 2025. doi: 10.1109/JSEN.2025.3557407.

C. Chi, J. Clapham, G. Blackwell, L. Cloud, I. Pretzer-Aboff, H. Shao, and G. Zhou. FM-FoG: A real-time foundation model-based wearable system for freezing-of-gait mitigation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(3), September 2026. doi: 10.1145/3831651.

Seffi Cohen, Nurit Cohen-Inger, and Lior Rokach. Bagstacking: An integrated ensemble learning approach for freezing of gait detection in parkinson's disease. *Information*, 15(12):822, December 2024. ISSN 2078-2489. doi: 10.3390/info15120822. URL http://dx.doi.org/10.3390/info15120822.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

J. D. Delgado-Terán, K. Hilbrants, D. Mahmutović, A. L. Silva de Lima, R. J. A. Wezel, and T. Heida. Ankle Sensor-Based Detection of Freezing of Gait in Parkinson's Disease in Semi-Free Living Environments. *Sensors*, 25(6):1895, 2025. doi: 10.3390/s25061895.

Hongfei Du, Jiacheng Shi, Sidi Lu, Gang Zhou, and Ye Gao. Sparse autoencoders for interpretable emotion control in text-to-speech, 2026. URL https://arxiv.org/abs/2606.01479.

Benjamin Filtjens, Pieter Ginis, Alice Nieuwboer, Muhammad Raheel Afzal, Joke Spildooren, Bart Vanrumste, and Peter Slaets. Modelling and identification of characteristic kinematic features preceding freezing of gait with convolutional neural networks and layer-wise relevance propagation. *BMC Medical Informatics and Decision Making*, 21(1):341, 2021. doi: 10.1186/s12911-021-01699-0.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Debin Huang, Chan Wu, Yiwen Wang, Zheyuan Zhang, Cheng Chen, Li Li, Wei Zhang, Zixuan Zhang, Jinyu Li, Yuzhu Guo, and Guiyun Cui. Episode-level prediction of freezing of gait based on wearable inertial signals using a deep neural network model. *Biomedical Signal Processing and Control*, 88:105613, 2024. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2023.105613. URL https://www.sciencedirect.com/science/article/pii/S1746809423010467.

Matīss Kalnāre, Sofoklis Kitharidis, Thomas Bäck, and Niki van Stein. Mechanistic interpretability for transformer-based time series classification, 2025. URL https://arxiv.org/abs/2511.21514.

K. Koltermann, J. Clapham, G. Blackwell, W. Jung, E. N. Burnet, Y. Gao, H. Shao, L. Cloud, I. Pretzer-Aboff, and G. Zhou. Gait-guard: Turn-aware freezing of gait detection for non-intrusive intervention systems. In *IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 61–72, 2024. doi: 10.1109/chase60773.2024.00016.

Nikita Koriagin, Georgii Aparin, Nikita Balagansky, and Daniil Gavrilov. Interpreting and steering a text-to-speech language model with sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.10029.

Yuri Kwon, Sang Hoon Park, Ji-Won Kim, Yeji Ho, Hyeong-Min Jeon, Min-Ju Bang, Gu-In Jung, Seon-Min Lee, Gwang-Moon Eom, Seong-Beom Koh, Jeong-Whan Lee, and Hyeong Seok Jeon. A practical method for the detection of freezing of gait in patients with Parkinson's disease. *Clinical Interventions in Aging*, 9:1709–1719, October 2014. doi: 10.2147/CIA.S69773.

William Lehn-Schiøler, Magnus Ruud Kjær, Rahul Thapa, Magnus Guldberg Pedersen, Anton Mosquera Storgaard, Nick Williams, Radu Gatej, Tue Lehn-Schiøler, Andreas Brink-Kjær, Sadasivan Puthusserypady, Sándor Beniczky, James Zou, and Lars Kai Hansen. Mechanistic interpretability of eeg foundation models via sparse autoencoders, 2026. URL https://arxiv.org/abs/2605.13930.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation, 2025. URL https://arxiv.org/abs/2412.05276.

S. R. Lord, H. Bindels, M. Ketheeswaran, M. A. Brodie, A. D. Lawrence, J. C. T. Close, A. L. Whone, Y. Ben-Shlomo, and E. J. Henderson. Freezing of gait in people with Parkinson's disease: Nature, occurrence, and risk factors. *Journal of Parkinson's Disease*, 10(2):631–640, 2020. doi: 10.3233/JPD-191813.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014. URL https://arxiv.org/abs/1312.5663.

Anurag Mishra. Dissecting chronos: Sparse autoencoders reveal causal feature hierarchies in time series foundation models, 2026. URL https://arxiv.org/abs/2603.10071.

S. T. Moore, H. G. MacDougall, and W. G. Ondo. Ambulatory monitoring of freezing of gait in Parkinson's disease. *Journal of Neuroscience Methods*, 167(2):340–348, 2008. doi: 10.1016/j.jneumeth.2007.08.023.

John G Nutt, Bastiaan R Bloem, Nir Giladi, Mark Hallett, Fay B Horak, and Alice Nieuwboer. Freezing of gait: moving forward on a mysterious clinical phenomenon. *The Lancet Neurology*, 10(8):734–744, 2011. doi: 10.1016/S1474-4422(11)70143-0.

Khalid Oublal, Quentin Bouniot, Qi Gan, Stephan Clémençon, and Zeynep Akata. Timesae: Causal sparse decoding for faithful explanations of black-box time series models, 2026. URL https://arxiv.org/abs/2601.09776.

A. Salomon, E. Gazit, P. Ginis, et al. A machine learning contest enhances automated freezing of gait detection and reveals time-of-day effects. *Nature Communications*, 15:4853, 2024. doi: 10.1038/s41467-024-49027-0.

J. D. Schaafsma, Y. Balash, T. Gurevich, A. L. Bartels, J. M. Hausdorff, and N. Giladi. Characterization of freezing of gait subtypes and the response of each to levodopa in parkinson's disease. *European Journal of Neurology*, 10 (4):391–398, 2003. doi: https://doi.org/10.1046/j.1468-1331.2003.00611.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1468-1331.2003.00611.x.

Boyan Wang, Xuegang Hu, Rongjun Ge, Chenchu Xu, Jinglin Zhang, Zhifan Gao, Shu Zhao, and Kemal Polat. Prediction of freezing of gait in parkinson's disease based on multi-channel time-series neural network. *Artificial Intelligence in Medicine*, 154: 102932, 2024. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2024.102932. URL https://www.sciencedirect.com/science/article/pii/S093336572400174X.

Yi Xia, Hua Sun, Baifu Zhang, Yangyang Xu, and Qiang Ye. Prediction of freezing of gait based on self-supervised pretraining via contrastive learning. *Biomedical Signal Processing and Control*, 89:105765, 2024. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2023.105765. URL https://www.sciencedirect.com/science/article/pii/S1746809423011989.

A Model Implementation Details

Table 2: **Model Implementation details.**

Component	Setting
Encoder and pretraining	
Input channels	12 bilateral IMU channels
Sampling rate	100 Hz
Temporal lengths	128 / 256 / 512 samples
Input projection	$12 \to 128$
Hidden dimension	128
Transformer blocks	6
Attention heads	8 (head dimension 16)
Feed-forward network	$128 \to 512 \to 128$
FFN activation	ReLU
Block structure	Residual connections + LayerNorm
Masked fraction	30%
Reconstruction head	$128 \to 12$
Pretraining loss	MSE on masked positions
Pooling	Mean pooling
Cross-scale fusion	Concatenate pooled representations
Classifier	$384 \to 128 \to 2$
Classifier dropout	0.4
Optimizer	AdamW with cosine annealing
Sparse autoencoder	
SAE dimensions	$128 \to 2048 \to 128$
TopK sparsity	$k = 32$
Training objective	Reconstruction MSE
Additional ℓ_1 penalty	None
Optimizer	AdamW
Learning rate	3×10^{-4}
Validation split	5%
Decoder normalization	Unit-norm during training
Inactive-feature handling	Periodic resampling
Alive-feature threshold	Firing rate $> 10^{-3}$

This section records implementation details.

Multi-scale encoder and pretraining. Each timestep contains 12 bilateral IMU channels: accelerometer x, y, z followed by gyroscope x, y, z for each foot. The three encoder streams operate on 128, 256, and 512 samples at 100 Hz, corresponding to 1.28, 2.56, and 5.12 s. The streams use the same configuration but have independent parameters. Unlike the pretraining model from which the encoder design was adapted, the current multi-scale model does not use a sensor-location token or location embedding.

Masked-sequence pretraining randomly removes 30% of sensor readings and reconstructs the original IMU values at the masked positions. A linear projection maps each 12-dimensional bilateral timestep to the 128-dimensional model space. Six Transformer blocks, each with eight self-attention heads and a $128 \to 512 \to 128$ feed-forward network, process the sequence before a linear reconstruction head maps tokens back to the 12-channel sensor space. The three temporal encoders use this same configuration but are trained as separate parameter sets.

FoG detection uses focal loss to address class imbalance. Batch size, learning rate, and other hyperparameters are tuned on the validation set. Training is performed on an NVIDIA A6000 GPU.

SAE training. For SAE training, tokens from all three frozen encoder streams are pooled into one training population. The SAE uses a 2048-feature dictionary with TopK sparsity $k = 32$. Decoder columns are normalized during training. After 20 epochs, 284 features satisfy the firing-rate criterion, and steering identifies 38 as model-relevant under $\Delta_f \geq 0.10$. All 38 steering-selected features were inspected under the same characterization procedure; nine exhibited a consistent recur-

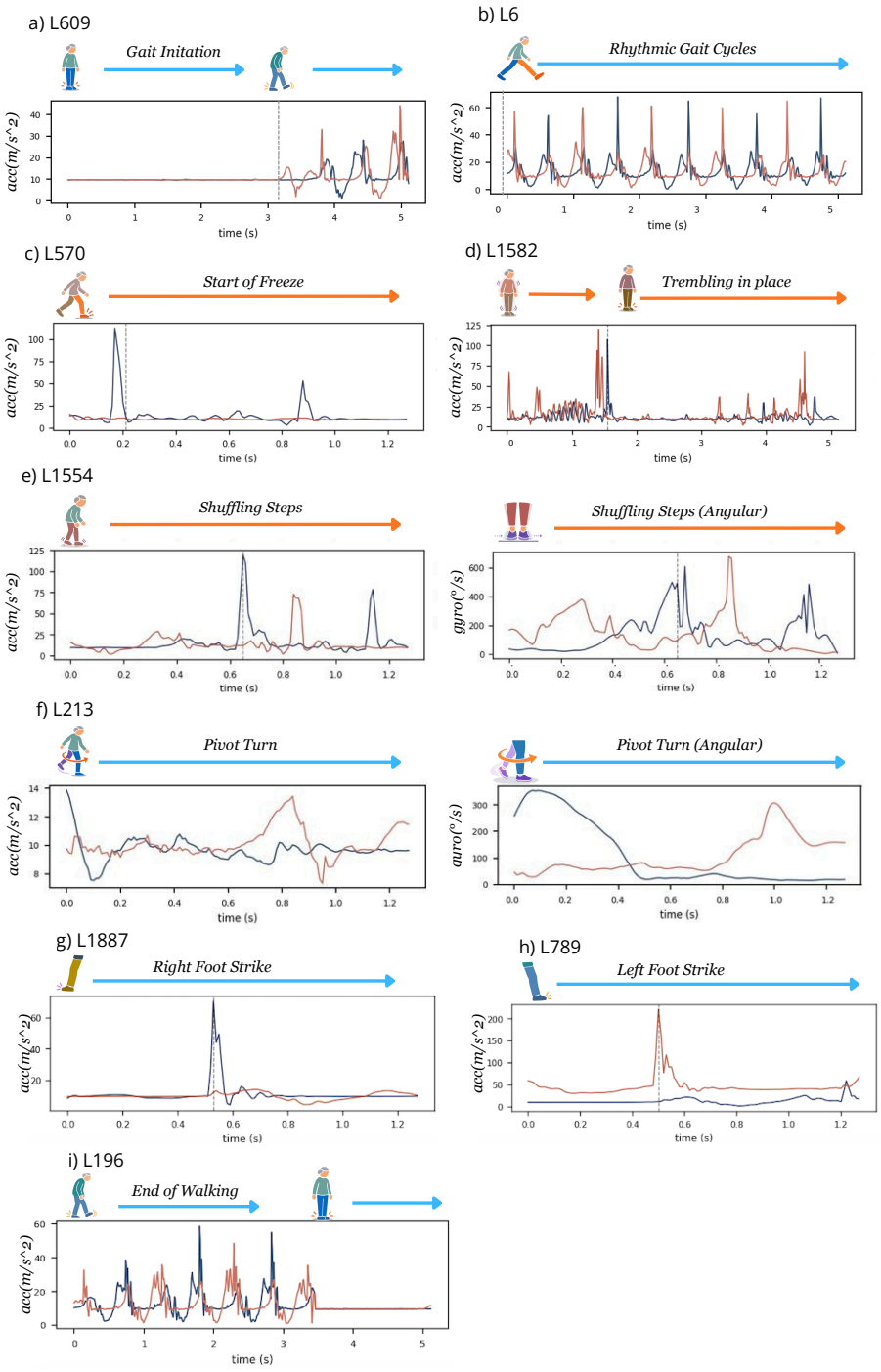


Figure 5: **Complete atlas of the nine discovered movement concepts.** Each panel shows a representative high-activation IMU window. The concepts span gait initiation, bilateral foot strikes, rhythmic gait cycles, start of freeze, trembling in place, shuffling steps, pivot turn, and end of walking. Acceleration is reported in $\mathrm{m/s^2}$ and angular velocity in $^\circ/\mathrm{s}$; red and blue denote the left and right foot, respectively.

ring movement pattern across de-duplicated high-activation windows and were assigned movement-concept interpretations.

Table 3: **Summary of the nine discovered movement concepts.** ↑ indicates that positive steering increases $P(\mathrm{FoG})$; ↓ indicates that it decreases $P(\mathrm{FoG})$.

Feature	Movement concept	Dominant scale	Steering
L609	Gait Initiation	5.12 s	↓
L789	Left Foot Strike	1.28 s	↓
L1887	Right Foot Strike	1.28 s	↓
L6	Rhythmic Gait Cycles	5.12 s	↓
L570	Start of Freeze	1.28 s	↑
L1582	Trembling in Place	5.12 s	↑
L1554	Shuffling Steps	1.28 s	↑
L213	Pivot Turn	1.28 s	↓
L196	End of Walking	5.12 s	↓

B Complete Movement-Concept Atlas

Figure 5 collects representative high-activation windows for all nine final movement concepts, including the three concepts not analyzed in detail in the main text: left-foot strike, right-foot strike, and end of walking. Examples are selected using the window-level activation defined in Eq. 5. The complete set shows that the recovered features span localized unilateral events, bilateral gait organization, transitions between locomotor states, FoG-specific and PD-associated gait patterns, and coordinated rotational actions.

Table 3 summarizes the final concept set together with its dominant temporal scale and functional steering direction. The two single-foot-strike features form a laterality-specific pair at the same short temporal scale, while gait initiation and end of walking form complementary long-context transitions. Positive steering of Start of Freeze, Trembling in Place, and Shuffling Steps raises $P(\mathrm{FoG})$, indicating that the model treats both FoG-specific patterns and associated PD gait abnormalities as positive evidence for FoG.

C Full Steering Effects

Complete steering responses. Using the additive intervention defined in Eq. 3 and the steering settings in the main paper, Figure 6 reports the complete responses for all nine final concepts. The six concepts shown in the main paper are supplemented here with L789, L1887, and L196. Positive steering of L570, L1582, and L1554 raises the FoG prediction, whereas positive steering of L609, L789, L1887, L6, L213, and L196 lowers it. Thus, Start of Freeze, Trembling in Place, and Shuffling Steps share a positive functional direction, although they span both FoG-specific phenomena and associated PD gait abnormalities, while the remaining movement concepts provide negative evidence for FoG under the tested intervention.

The additional single-foot concepts provide a useful consistency check. Strengthening either the left-foot-strike or right-foot-strike feature lowers the FoG prediction, despite the features encoding opposite sides of the body. End of Walking is also negative-direction, indicating that its functional role is distinct from the FoG-transition feature L570 even though both involve a change in locomotor state.

Temporal-scale preference. Scale identity is not provided to the shared SAE. We therefore quantify temporal preference only after training. For feature f and encoder scale s, the scale-normalized firing rate is

$$r_{f,s} = \frac{1}{N_s} \sum_{i:s_i=s} \mathbf{1}[z_f(h_i) > 0], \tag{6}$$

where N_s is the number of evaluated tokens originating from scale s. Normalizing by N_s prevents the different sequence lengths from mechanically determining the firing rate. The dominant scale is

$$s_f^\star = \underset{s\in\{1.28,\,2.56,\,5.12\}}{\arg\max} r_{f,s}. \tag{7}$$

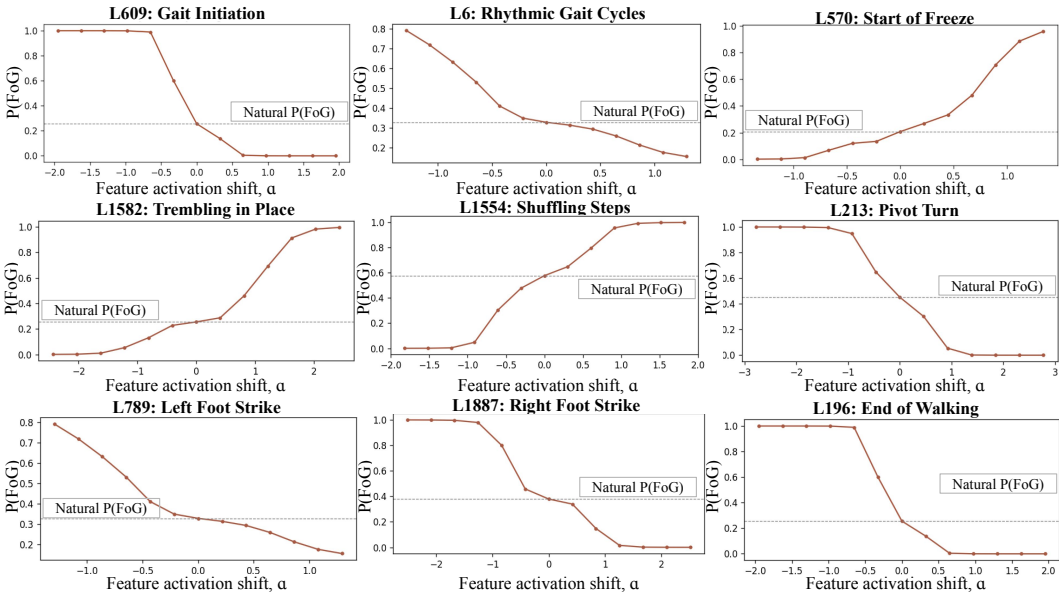


Figure 6: **Complete feature-steering responses for all nine movement concepts.** Each curve reports mean $P(\text{FoG})$ over a fixed feature-active host set as the representation is shifted along the corresponding unit-norm SAE decoder direction. Dashed lines denote the natural, unmodified prediction.

Five of the nine concepts prefer the 1.28 s stream and four prefer the 5.12 s stream; none is dominated by the 2.56 s stream (Table 3). The short-context features include both unilateral foot strikes, Start of Freeze, Shuffling Steps, and Pivot Turn, whose defining structure is visible within a compact local interval. The long-context features are Gait Initiation, Rhythmic Gait Cycles, Trembling in Place, and End of Walking. For these concepts, interpretation depends on extended before/after context, repeated gait cycles, or sustained movement organization. The scale pattern therefore reflects the temporal extent needed to distinguish different movement patterns rather than a simple "transition versus state" split.

C.1 Random-Direction Intervention Ablation

The steering curves in Fig. 6 show that intervening along the learned SAE directions can substantially change the FoG prediction. To test whether these effects are specific to the learned feature directions, not a generic consequence of perturbing the hidden representation, we compare each SAE intervention against norm-matched random directions.

Random-direction baseline. For each movement-concept feature f, we sample 1,000 random unit vectors $r_j \in \mathbb{R}^{128}$, matching the normalization of the SAE decoder direction d_f. We repeat the intervention using the same feature-active host windows, the same naturally active token positions, and the same feature-specific steering range as in Fig. 6. Thus, only the intervention direction is randomized:

$$\tilde{h}^{(j)}_{s,t}(\alpha) = \begin{cases} h_{s,t} + \alpha r_j, & z_f(h_{s,t}) > 0, \\ h_{s,t}, & z_f(h_{s,t}) = 0. \end{cases} \tag{8}$$

For each random direction, we compute the same steering effect

$$\Delta^{(j)}_f = \max_{\alpha \in \mathcal{A}_f} \bar{p}^{(j)}_f(\alpha) - \min_{\alpha \in \mathcal{A}_f} \bar{p}^{(j)}_f(\alpha), \tag{9}$$

Table 4: **Random-direction intervention control.** Random values summarize the distribution of steering effects from 1,000 norm-matched random directions. The SAE comparison indicates whether the steering range shown in Fig. 6 exceeds the 95th percentile or the maximum random effect.

Feature	Movement concept	Random mean $\pm$ SD	Random 95th	SAE vs. random
L609	Gait Initiation	0.048 ± 0.028	0.085	$>$ max
L789	Left Foot Strike	0.066 ± 0.056	0.189	$>$ max
L1887	Right Foot Strike	0.109 ± 0.134	0.412	$>$ max
L6	Rhythmic Gait Cycles	0.196 ± 0.175	0.629	$>$ 95th percentile
L570	Start of Freeze	0.062 ± 0.059	0.200	$>$ max
L1582	Trembling in Place	0.133 ± 0.165	0.542	$>$ max
L1554	Shuffling Steps	0.166 ± 0.188	0.640	$>$ max
L213	Pivot Turn	0.062 ± 0.056	0.182	$>$ max
L196	End of Walking	0.022 ± 0.039	0.093	$>$ max

where $\bar{p}_f^{(j)}(\alpha)$ is the mean FoG probability over the fixed host set after steering along r_j. We compare this null distribution with the intervention range produced by the learned SAE direction shown in Fig. 6.

Results. The learned SAE directions produce substantially larger intervention effects than random directions in the same representation space (Table 4). For eight of the nine movement concepts, the steering range shown in Fig. 6 exceeds the largest effect observed among 1,000 random directions. Rhythmic Gait Cycles (L6) is the only exception: its learned-direction effect exceeds the 95th percentile of the random distribution but not its maximum.

These results show that the large changes in FoG probability observed under feature steering are not explained by arbitrary perturbations of comparable magnitude. Rather, the effects are specific to the learned SAE directions, providing additional evidence that the identified features are functionally linked to the model's FoG decisions.

D Additional Movement Concepts Validation Analysis

This section provides the complete construction of the cross-subject validation tests summarized in Sec. 3.2. Movement characterization uses 10 participants, while evaluation uses a disjoint set of 13 participants. All probes and reference prototypes are constructed using only the characterization subjects and are applied to the evaluation subjects without refitting.

Raw-Signal Representation and Activation Labels. Each validation sample is one IMU window w. To evaluate the discovered patterns independently of the SAE representation, we convert each window into a 128-dimensional raw-signal descriptor. Let $\boldsymbol{a}_L(t), \boldsymbol{a}_R(t) \in \mathbb{R}^3$ denote left- and right-foot acceleration and $\boldsymbol{\omega}_L(t), \boldsymbol{\omega}_R(t) \in \mathbb{R}^3$ denote angular velocity. We compute the magnitude of each signal, resample each trajectory to 32 points, and concatenate them:

$$\begin{aligned}\boldsymbol{\phi}(w) = \text{concat}\Big(&\mathcal{R}_{32}(\|a_L(t)\|_2), \mathcal{R}_{32}(\|\omega_L(t)\|_2),\\ &\mathcal{R}_{32}(\|a_R(t)\|_2), \mathcal{R}_{32}(\|\omega_R(t)\|_2)\Big) \in \mathbb{R}^{128},\end{aligned} \tag{10}$$

where $\mathcal{R}_{32}$ denotes temporal resampling to 32 points.

For feature f, the window-level activation is defined in Eq. 5. We assign the binary activation label

$$y_f(w) = \mathbf{1}[z_f(w) > 0], \tag{11}$$

where $y_f = 1$ denotes a feature-active window and $y_f = 0$ an inactive window. Inactive controls are sampled to match the temporal-scale distribution of the active windows. Controls are not additionally matched by subject.

Recurrence. Recurrence measures whether the raw-signal structure associated with feature activation remains detectable in subjects not used for characterization. For each feature f, we fit a logistic-regression probe on characterization windows using $\boldsymbol{\phi}(w)$ as input and $y_f(w)$ as the binary target:

$$q_f(w) = P(y_f = 1 \mid \boldsymbol{\phi}(w)). \tag{12}$$

The fixed probe is then applied to evaluation-subject windows, producing one continuous active-class score $q_f(w)$ for each window. Recurrence ROC-AUC is computed from the held-out binary labels $\{y_f(w_i)\}$ and corresponding probe scores $\{q_f(w_i)\}$. A high recurrence AUC therefore indicates that active windows in unseen subjects retain raw-signal structure that distinguishes them from scale-matched inactive controls.

Prototype Consistency. Recurrence requires active windows to be separable from inactive controls but does not require them to retain the characteristic signal form observed during characterization. We therefore construct a fixed prototype for each feature from its active characterization windows:

$$\boldsymbol{\mu}_f = \frac{1}{|\mathcal{W}_f^{\text{char}}|} \sum_{w \in \mathcal{W}_f^{\text{char}}} \boldsymbol{\phi}(w), \tag{13}$$

where $\mathcal{W}_f^{\text{char}}$ is the set of active characterization windows for feature f.

No classifier is trained for this test. Each evaluation window receives one scalar score given by its cosine similarity to the fixed prototype:

$$s_f(w) = \frac{\boldsymbol{\phi}(w)^\top \boldsymbol{\mu}_f}{\|\boldsymbol{\phi}(w)\|_2 \|\boldsymbol{\mu}_f\|_2}. \tag{14}$$

Prototype-consistency ROC-AUC is computed from the same active/inactive labels $\{y_f(w_i)\}$ and the similarity scores $\{s_f(w_i)\}$. A high AUC therefore means that active windows from unseen subjects are consistently more similar to the characterization-subject prototype than matched inactive controls.

Pairwise Specificity. Specificity tests whether different discovered features correspond to distinguishable raw-signal patterns. For each pair of features (f_i, f_j), we collect their active characterization windows and assign the binary feature-identity target

$$y_{ij}(w) = \begin{cases} 1, & w \text{ is assigned to } f_i, \\ 0, & w \text{ is assigned to } f_j. \end{cases} \tag{15}$$

If a window activates both features, it is assigned to the feature with the larger window-level activation $z_f(w)$. A logistic-regression probe takes $\boldsymbol{\phi}(w)$ as input and predicts this feature identity. The fixed probe is then applied to the disjoint evaluation subjects, producing one feature-identity score per window. Pairwise ROC-AUC is computed from the held-out feature labels and classifier scores. High specificity AUC therefore indicates that the two feature-associated movement patterns remain distinguishable across subjects.

Per-Concept Cross-Subject Results. Table 5 reports the complete recurrence, prototype-consistency, and coverage results. Recurrence AUC ranges from 0.90 to 0.99 (median 0.96), while prototype-consistency AUC ranges from 0.88 to 1.00 (median 0.95). The nine concepts occur in 15–23 participants (median 19), indicating that the identified patterns are not confined to isolated individuals.

The two unilateral foot-strike concepts achieve particularly high recurrence and prototype consistency, indicating that their feature-associated signal structure transfers strongly across participants. Start of Freeze has the lowest recurrence and prototype AUCs among the nine concepts (0.90 and 0.88, respectively), but remains well above chance in both tests. This variation suggests that some movement concepts have more stereotyped raw-signal realizations than others.

Pairwise Concept Relationships. Table 6 reports held-out specificity AUC for all 36 concept pairs. The overall median is 0.875. Same-scale pairs have a median AUC of 0.915, compared with

Table 5: **Cross-subject validation of the nine discovered movement concepts.** Recurrence AUC measures held-out separation between feature-active windows and scale-matched inactive controls using a logistic probe. Prototype AUC measures whether held-out active windows are more similar than controls to a fixed prototype constructed from characterization subjects. Coverage denotes the number of participants in whom the feature occurs.

Feature	Movement concept	Recurrence AUC	Prototype AUC	Coverage
L609	Gait Initiation	0.96	0.93	22
L789	Left Foot Strike	0.99	1.00	18
L1887	Right Foot Strike	0.98	0.98	19
L6	Rhythmic Gait Cycles	0.97	0.97	23
L570	Start of Freeze	0.90	0.88	15
L1582	Trembling in Place	0.97	0.96	18
L1554	Shuffling Steps	0.94	0.91	16
L213	Pivot Turn	0.95	0.94	20
L196	End of Walking	0.96	0.95	21

Table 6: **Held-out pairwise specificity AUC among the nine movement concepts.** Each entry measures how well a logistic probe distinguishes active windows of the corresponding feature pair in the disjoint evaluation subjects. Windows activating both features are assigned to the feature with the stronger window-level activation.

	L609	L789	L1887	L6	L570	L1582	L1554	L213	L196
L609	–	0.82	0.80	0.78	0.88	0.99	0.84	0.86	0.74
L789	0.82	–	1.00	0.84	0.90	0.94	0.92	0.97	0.80
L1887	0.80	1.00	–	0.85	0.91	0.95	0.93	1.00	0.81
L6	0.78	0.84	0.85	–	0.93	0.89	0.82	0.87	0.76
L570	0.88	0.90	0.91	0.93	–	0.86	0.72	0.98	0.85
L1582	0.99	0.94	0.95	0.89	0.86	–	0.70	0.90	1.00
L1554	0.84	0.92	0.93	0.82	0.72	0.70	–	0.88	0.86
L213	0.86	0.97	1.00	0.87	0.98	0.90	0.88	–	0.89
L196	0.74	0.80	0.81	0.76	0.85	1.00	0.86	0.89	–

0.855 for cross-scale pairs. Because concepts from the same encoder stream share the same temporal context, their stronger separation indicates that specificity is not explained simply by scale identity.

The pairwise structure provides additional evidence that the recovered features encode specific movement organization rather than only coarse temporal context. Left and Right Foot Strike are perfectly separable (AUC $= 1.00$) despite representing the same type of localized event at the same 1.28 s scale. Their separation therefore reflects laterality in the underlying movement rather than temporal-scale differences. Other highly separable same-scale pairs include Right Foot Strike versus Pivot Turn (AUC $= 1.00$) and Start of Freeze versus Pivot Turn (AUC $= 0.98$).

Conversely, several of the least-separable pairs correspond to movements with related signal structure. Trembling in Place versus Shuffling Steps has AUC 0.70, and Start of Freeze versus Shuffling Steps has AUC 0.72; all three capture disrupted gait patterns relevant to FoG detection. Gait Initiation versus End of Walking has AUC 0.74, while Rhythmic Gait Cycles versus End of Walking has AUC 0.76, consistent with their shared locomotor structure. The specificity matrix therefore reflects graded relationships among movement concepts rather than arbitrary separability of feature identities.

E SAE Information Loss and Configuration Sensitivity

Representation fidelity. Before interpreting the learned sparse features, we verify that the SAE preserves the encoder representation and the downstream FoG prediction. For a validation token h and its SAE reconstruction $\hat{h}$, we measure variance explained as

$$\mathrm{VE} = 1 - \frac{\mathbb{E}\left[\|\boldsymbol{h} - \hat{\boldsymbol{h}}\|_2^2\right]}{\mathbb{E}[\|\boldsymbol{h} - \boldsymbol{\mu}_h\|_2^2]}, \tag{16}$$

Table 7: **Sensitivity to SAE configuration.** R^2 measures validation reconstruction fidelity. Alive features satisfy the firing-rate criterion $r_f > 10^{-3}$.

d_{SAE}	k	Validation R^2	Alive features
512	32	0.9973	245 / 512
1024	16	0.9931	198 / 1024
1024	32	0.9974	258 / 1024
2048	16	0.9923	194 / 2048
2048	**32**	**0.9976**	**284 / 2048**

where μ_h is the mean validation token representation. The SAE used in the main analysis, with $d_{\mathrm{SAE}} = 2048$ and TopK sparsity $k = 32$, the SAE achieves a validation reconstruction R^2 of 0.9976.

We additionally measure whether this reconstruction preserves the downstream computation performed by the frozen FoG model. We replace the original encoder tokens with their SAE reconstructions and propagate them through the unchanged pooling, fusion, and classification layers. The original and reconstructed FoG probabilities have Pearson correlation $r = 0.9999$, with a mean absolute probability difference of 0.0013. FoG-detection F1 changes by approximately -0.0001. Thus, although the sparse reconstruction is not exact in representation space, it preserves nearly all task-relevant information at the analyzed layer.

Sensitivity to SAE configuration. We examine the sensitivity of the decomposition to dictionary width d_{SAE} and TopK sparsity k. We train SAEs with $d_{\mathrm{SAE}} \in \{512, 1024, 2048\}$ and $k \in \{16, 32\}$ where available, using the same training procedure. Table 7 reports validation reconstruction R^2 and the number of alive features, where a feature is considered alive if its firing rate exceeds 10^{-3}.

At matched dictionary widths, $k = 32$ consistently provides both higher reconstruction fidelity and a larger set of active features than $k = 16$. For $d_{\mathrm{SAE}} = 1024$, increasing k from 16 to 32 raises R^2 from 0.9931 to 0.9974 and the number of alive features from 198 to 258. The same pattern holds at $d_{\mathrm{SAE}} = 2048$, where R^2 increases from 0.9923 to 0.9976 and the number of alive features from 194 to 284. We therefore use $k = 32$ for the main analysis.

Among the $k = 32$ configurations, reconstruction fidelity is already near saturation: validation R^2 changes only from 0.9973 at $d_{\mathrm{SAE}} = 512$ to 0.9974 at 1024 and 0.9976 at 2048. Dictionary width therefore has little effect on information preservation in this range. Instead, increasing the width provides additional capacity for separating recurring representation patterns. During inspection of the 512-dimensional SAE, we observed greater qualitative redundancy, with multiple features responding to similar movement patterns, suggesting that the narrower dictionary can allocate multiple features to closely related structure.

We use $d_{\mathrm{SAE}} = 2048$, $k = 32$ as a high-fidelity, overcomplete operating point for the movement-concept analysis. This setting retains the near-lossless reconstruction observed at smaller widths while providing greater dictionary capacity and the largest number of alive features among the tested configurations. Importantly, our choice is not based on a claim that $d_{\mathrm{SAE}} = 2048$ is uniquely optimal: the 1024- and 2048-dimensional SAEs have nearly identical reconstruction fidelity. Rather, the wider dictionary is used to reduce capacity constraints on the feature decomposition while preserving downstream model behavior.

The qualitative redundancy observation was not converted into a systematic semantic-overlap metric across all SAE configurations. We therefore interpret this experiment as a configuration-sensitivity analysis rather than evidence for a unique optimal dictionary size. The main analysis only requires that the selected SAE provide a sparse, high-fidelity decomposition from which functionally relevant and reproducible movement concepts can be identified.